Open Data in the Earth and Climate Sciences

Sarah Callaghan
British Atmospheric Data Centre, UK

Introduction

It is commonly acknowledged that data is the foundation of science—without access to the data used to derive results and conclusions it is not possible for other researchers to verify and reproduce the science. Reproducibility, though a fundamental part of the scientific process, is a difficult principle to follow for a number of reasons. This is especially true in the Earth and climate sciences, where even a simple experiment of taking an outdoor air temperature measurement may vary from one minute to the next, with no possibility of repeating measurements that occurred in the past.

How to cite this book chapter:
Callaghan, S. 2014. Open Data in the Earth and Climate Sciences. In: Moore, S. A. (ed.) *Issues in Open Research Data*. Pp. 89–106. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/ban.f

Access to and openness of data will facilitate reproducibility of science in the future. In the present, access to data encourages increased collaboration and reuse, allowing the identification of new multidisciplinary research avenues.

Along with the principle of reproducibility, openness of data in the Earth sciences allows for a better understanding of vital systems, including climate and weather. It is simply not possible for researchers to take measurements of every meteorological parameter at every point on the surface of the Earth. Past weather measurements, such as those found in historical ships logs (Oliver & Kington 1970; Garcia-Herrera et al. 2005; Chappell & Lorrey 2013), are invaluable for filling in the gaps in our understanding of climate change.

The Challenges of Earth and Climate Science Data

The majority of Earth science data is observational, which means that it is irreproducible. Without the aid of a time machine, it is simply not possible to travel back to last week to take a measurement that was forgotten at the time. For the same reason, we need to manage and archive the data that was collected last week, because if it is lost, it is gone for good. This is particularly relevant for fast-changing phenomenon such as weather, whereas the timescales for measurement are a bit more forgiving when it comes to the geological sciences (though not always—see for example the differences in measurements of Mount St Helens mere minutes before and after its eruption (US Geological Survey 2000)).

By contrast, much climate science is done using large and complicated software models to simulate the climate. In theory, because these are computer models, the results are reproducible

by simply re-running the model with the same input parameters. In practice, however, this is not possible due to the complexity of the models, and a lack of standardisation of the metadata required to initiate them and reproduce model runs. The recent European Union Framework 7 project Metafor (Guilyardi et al. 2013) attempted to standardise and collect the metadata needed for the climate model runs done as part of the Fifth Climate Intercomparison Project (CMIP5), which fed into the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Metafor used a web-based questionnaire-type system, with associated controlled vocabulary, which took climate-modelling centres approximately two weeks to fill in—a not inconsiderable effort!

Earth science data also comes in a wide variety of formats (almost one for each type of measuring instrument used), and the datasets produced can get up to terabytes in size, as well as taking months, years, or even decades to complete.

As an (incomplete) example, the UK's Natural Environment Research Council (NERC) funds seven data centres that between them have responsibility for the long-term management of NERC's environmental data holdings. These data centres deal with a variety of environmental measurements, along with the results of model simulations in atmospheric science; Earth sciences; Earth observation; marine science; polar science; terrestrial and freshwater science, hydrology and bioinformatics; and solar-terrestrial physics and space weather.

The NERC environmental data centres hold many different types of datasets, including time series, with some series some being continually updated (e.g. meteorological measurements); large four-dimensional synthesised datasets (e.g. climate, oceanographic, hydrological, and numerical weather prediction model

data generated on a supercomputer); two-dimensional scans (e.g. satellite data, weather radar data); two-dimensional snapshots (e.g. cloud cameras); traces through a changing medium (e.g. radiosonde launches, aircraft flights, ocean salinity and temperature); datasets consisting of data from multiple instruments as part of the same measurement campaign; and physical samples (e.g. fossils).

Data is also produced in a variety of ways by a variety of researchers, ranging from individual researchers, to small research groups, up to entire institutions.

Large research groups and institutions tend to have a more 'industrial' process for developing the data, where standards for data formats and metadata are well defined and adhered to by all participants. Openness of the data within the collaboration or project group is the norm, and systems are set in place to share the data within that group. The standardised data formats and metadata are a boon to helping the project members share data within their group, and would be useful for researchers using the data outside the group too. Often, however, the data are closed to all but the members of the group. Paradoxically, putting access restrictions in place on a collaborative workspace may make researchers more likely to open their data within that workspace and begin the process of sharing.

Small research groups are less likely to have standardised formats for data and/or metadata (unless they are part of a larger community, such as the atmospheric sciences where standardised file formats such as NetCDF are common). This does not make them any less open to sharing their data, but it does introduce an extra overhead of effort for the person being shared with, as they then have to learn the format and decipher the metadata (if any) before they can use the dataset.

Drivers for Openness

Measuring Earth science phenomena is expensive, often requiring expensive equipment such as ships or aircraft, or large networks of instruments such as rain gauges or radars. Funders are keen to ensure that the data collected as a result of their funding is archived and managed properly, not only to ensure the quality of the research, but also to enable reuse of the data by other researchers (both inside the domain of interest and related) thereby saving time, effort and money.

Members of the climate science community were pushed towards openness after the 'Climategate' affair, when, in November 2009, a server at the Climatic Research Unit (CRU) at the University of East Anglia was hacked and thousands of emails and computer files were copied to various locations on the Internet. This resulted in the spread of alleged malpractices found within the leaked CRU emails around the Internet, where it was claimed 'that these e-mails showed a deliberate and systematic attempt by leading climate scientists to manipulate climate data, arbitrarily adjusting and 'cherry-picking' data that supported their global warming claims and deleting adverse data that questioned their theories.' (House of Commons Science and Technology Committee 2010). In the resulting investigations, no evidence of fraud or scientific misconduct was found, and recommendations were made to avoid any such allegations in the future by opening up access to their supporting data.

Openness of data encourages reuse, and adds value to other research. One example is of a researcher using rainfall data in her studies of newt populations (British Atmospheric Data Centre 2007); her access to this data added an extra dimension to her studies, allowing her to draw more complete conclusions. Without

access to the data, her research would have been the poorer, as she would not have been able to make the required measurements herself in the context of her own investigation.

Barriers to Openness

Simply opening up a dataset for use by others is not enough. It is very easy to stick some data files on a departmental or personal website, with file names that may be clear to the producer, but are opaque to everyone else. Once in the file (assuming they can open it in the first place), a potential user may have to figure out what the various variables actually mean, and then dig through other information (published in journal articles or not) before they can really make use of the data. Just because data is open does not mean it is usable. Similarly, just because a dataset is archived does not mean it will still be usable in 20 years' time, especially given the rate of change in commonly used file formats such as Excel.

In an increasingly competitive environment for research funding, access to important datasets may be the only factor determining whether a grant is won or not. For this reason, there is a tendency for researchers to hoard data until they have extracted all the possible research benefit out of it. This can be combated by the research funders' policies on open data and embargoes.

In the absence of common practices or standards, some researchers have misgivings about making their data either freely or openly available, as they fear that other researchers may find errors or 'misuse the data', or that the researcher themselves will get 'scooped' and lose out on research funding (RIN 2008).

A Tale of Two Datasets: The Author's Personal Experience of Open Data

The datasets

Upon finishing her first degree, the author was hired by the Radio Communications Research Unit (RCRU, now the Chilbolton Group), at Science and Technology Facilities Council Rutherford Appleton Laboratory, UK, to process and analyse data received from ITALSAT, a communications satellite. The group was investigating the effects of rain, clouds and atmospheric gases on the received signal levels from radio beacons aboard geosynchronous satellites. Their aim was to determine the best way of counteracting the signal fading experienced by radio frequencies above 10 GHz when a rain storm blocks the path between the satellite transmitter and the receiver in the ground station. To perform these measurements, the RCRU installed and operated a number of receivers at different locations in Hampshire, UK. **Table 1** gives information about the experiments, including the locations, the measurement periods and the primary publications. It is important to note the significant delay between the completion of the ITALSAT experiment and the primary publication from it. Also, the primary work of the staff involved in the ITALSAT and Global Broadcast Service (GBS) experiments was to run and manage long-term measurement campaigns, meaning that writing up the experiments for publication was often a lower priority.

Pre-processing the received signal levels was the author's main job for several years. The received signal levels had to be processed to remove the diurnal variability introduced as the satellite varied in its orbit because of its age and the lack of fuel available to make station keeping adjustments. This pre-processing

Experiment	ITALSAT	GBS (Global Broadcast Service)
Frequencies studied	49.5, 39.6 and 18.7 GHz	20.7 GHz
Receive sites	Sparsholt (51° 04' N, 01° 26' W)	Sparsholt (51° 04' N, 01° 26' W) Chilbolton (51° 08' N, 01° 26' W) Dundee (56.45811° N, 2.98053° W)
Measurement period start	April 1997	Chilbolton: August 2003 Sparsholt: October 2003 Dundee: February 2004
Measurement period end	January 2001	August 2006
Primary publication(s)	Ventouras et al. 2006	Callaghan et al. 2008 Callaghan et al. 2013

Table 1: Key characteristics of the ITALSAT and GBS datasets.

involved four major steps, four different computer programmes and 16 intermediate files for each day of measurements. Each month of pre-processed data represented somewhere between a couple of days' and a week's worth of effort. It was a job where attention to detail and scientific knowledge and data experience were important.

Sharing the data

The ITALSAT raw and processed data were stored on the RCRU's servers, with a backup on CD on a shelf in the author's office (where it still resides).

We were approached by other radio propagation research groups to share our data, and in some cases we did so. Because the data were in a non-standard format, this involved sharing the software we used and, occasionally, physically sitting with

the new users, explaining how it had been created and what the files meant.

The first article about the ITALSAT dataset was published in 2003 (Otung & Savvaris 2003), three years before the first publication from the researchers who produced the data. We were not part of the author list on the 2003 paper, though I believe we got a group acknowledgement.[1] There was also at least one other occasion where we 'shared' our data with other researchers, who then went on to receive further funding for work in the same subject area that did not include us.

An added complication was that this data was (in theory) commercially valuable and could have been sold to telecommunications companies. Hence, in a number of cases, sharing required the development of non-disclosure agreements, in consultation with our contracts department, which took a lot of time and effort.

Eventually, we just hoarded the data, which was not good for us, or for science! It was only after the group's funding was changed, and our new funders mandated that all the group's data should be deposited in the British Atmospheric Data Centre (BADC), that we moved away from keeping the data on private servers in non-standard formats.

Opening the data

Both the ITALSAT and GBS datasets have now been archived in the BADC and have been assigned digital object identifiers (DOIs) to enable formal data citation to occur (STFC 2009a, 2009b, 2009c, 2012a, 2012b, 2012c). It is worth noting that the

[1] Unfortunately I cannot check as the referenced paper is behind a paywall.

DOIs for the GBS dataset were only assigned in April 2011, and the ITALSAT data DOIs were assigned in 2012—a long time after the completion of the datasets and their primary publications. Even though the datasets are now citeable and discoverable in the BADC, they are still not completely open, as they can only be downloaded by registered BADC users. However, there are no restrictions on who can become a BADC user. Also, the Chilbolton Group would like to monitor the use of these data and require an acknowledgement of the data source if they are used in any publication.

Detailed project reports were written about both the ITALSAT and GBS experiments and provided to the funders of the experiments. These reports are a valuable resource because they are significantly longer and more detailed than the journal publications, but because they are grey literature, access to them is limited. For the GBS experiment, the report is marked as 'commercial in confidence' and therefore cannot be made public. For ITALSAT, the documentation has fallen foul of changes in word processing software and key figures in the document cannot be viewed on-screen. This just goes to show that data curation applies to supporting documentation as much as it does to the datasets themselves.

Publications and the datasets

Ventouras and colleagues (2006) do not make any statement on data availability or the location of the raw data. The article does include some of the derived data in the form of tables and figures of cumulative distribution functions, but there is a crucial disconnect between the paper and the dataset on which it bases its conclusions.

Similarly, for the GBS dataset, the Callaghan et al. (2008) paper does not include any figures or tables of the processed data, instead only presenting figures showing the curves resulting from the analysis. These authors do comment about the location of the underlying data: 'The database collected as part of the GBS experiment has been submitted to the International Telecommunications Union (ITU-R) Study Group 3 for inclusion into its databanks.' These databanks are available online but it is not clear where the GBS experiment data can be found within them.[2]

Note that for both experiments, the final step (archiving the data or publication in a data journal) took place some time after the experiment was officially concluded. This would not be possible for many research groups because the researcher who did the majority of the data processing and analysis is very likely to have left that research group (as a result of finishing their PhD or postdoc, or finding a position elsewhere once the project funding finished).

Encouraging Openness: Carrots and Sticks

As mentioned earlier, the scientific consensus is changing to the belief that openness should be the norm rather than the exception (Royal Society 2012). But in order to encourage the researcher producing the data to open it and, more to the point, open it in a way that is useful to other users, rewards and sanctions are needed. Steps have already been made, with many research funders publishing data policies (RCUK 2013a, b; European Commission 2013; NSF 2010) that outline their expectations of their funded

[2] http://www.itu.int/ITU-R/index.asp?category=study-groups&rlink=rsg3&lang=en

researchers. The methods for applying sanctions have yet to be applied, or even defined.

Focus in the UK and elsewhere has been on the rewards that researchers can obtain by making their data open and usable. Researchers are used to getting credit for publishing papers about their research in academic journals, hence this mechanism is used to provide credit for publishing data. The mechanisms for data citation and publication are still under development, but early indications are that they will act as an incentive and encourage openness of data. For example, a survey of atmospheric science researchers carried out at the UK's National Centre for Atmospheric Science Conference in Bristol on 8–10 December 2008 showed that 67% of the 85 respondents agreed that they are more likely to deposit their data in a data centre if they can obtain academic credit through a data journal (Callaghan et al. 2009).

Publishing a Dataset in an Academic Context

Going back to the case study above, the GBS dataset differs from the ITALSAT dataset (and many others) in that it has been formally published in a data journal (Callaghan et al. 2013).

A data journal is an online journal that specialises in the publication of scientific data in a way that includes scientific peer-review. Most data journals publish short data papers cross-linked to, and citing, datasets that have been deposited in approved data centres.

A data paper is a short article that describes a dataset, and provides details of the dataset's collection, processing, software, file formats etc., allowing the reader to understand the when, how and why data was collected and what the data product is. The data paper does not require novel analyses or ground-breaking

conclusions, instead the dataset is presented 'as is', allowing the publication of negative results.

Data journals support the development and enhancement of the scholarly record by providing a mechanism for:

- peer-reviewing datasets;
- publishing datasets quickly, as the data journal does not require analysis or novelty in the publication;
- providing attribution and credit for the data collectors who might not be involved with the analysis, and therefore would not be eligible for author credit for an analysis paper; and
- enabling the discovery and understanding of datasets, and providing assurance of their quality and provenance.

Data journals are becoming more prevalent in the scientific publishing ecosystem, signifying a recognition by publishers and funders that a mechanism for publishing data is required (and encouraging openness and access to data). For many researchers, who may be concerned that 'making their data open' is synonymous with 'giving it away and getting no credit', re-framing data sharing in the context of data citation and publication reassures them, and provides a structure and a framework that is well understood, where precedence and attribution are an accepted part of the publication and citation process.

There are many issues that need to be dealt with to ensure the smooth running of data journals, including (but not limited to) providing guidance to reviewers on how exactly to go about peer-reviewing a dataset, and how to certify that a data repository is suitably trustworthy for hosting published data. Data journals also rely significantly on a linking mechanism that is robust and reliable to link the article to the dataset, especially in those cases where the dataset is archived in a repository outside of the journal

publisher's control. Linking between digital objects is commonplace on the Internet, but for the scholarly record to be maintained, the links between articles and datasets must be held to a higher standard of stability and reliability. These issues are not solved as of the date of this chapter, though there is a sizeable (and growing) community of researchers, librarians, data centre managers, academic publishers and research funders who are coming together to propose solutions and guidance for these problems.

Conclusions

Changing scientific culture is difficult and requires both incentives and disincentives, along with systems put in place to ease the process of change, and a critical mass of researchers who wish to make the change. The Earth and climate sciences have experienced their share of issues with lack of openness in the past (on a national level with Climategate, and on a multitude of personal levels, one example of which as described in this chapter). However, the push on researchers is definitely towards openness, and research funders are putting policies in place to support this. Bringing data into the academic publication process is potentially a very valuable way to encourage researchers to be more open with their data, while providing them with the credit they deserve for doing so.

References

British Atmospheric Data Centre 2007 National Database Of Atmospheric and Weather Data Tops 10,000 Users [British Academic Data Centre news release], 3 September. Available at https://badc.nerc.ac.uk/community/news/070906_Press.html [Last accessed 11 August 2014].

Callaghan, S A, Boyes, B, Couchman, A, Waight, J, Walden, C J and Ventouras, S 2008 An investigation of site diversity and comparison with itu-r recommendations. *Radio Science*, 43: RS4010. DOI:10.1029/2007RS003793.

Callaghan, S, Hewer, F, Pepler, S, Hardaker, P and Gadian, A 2009 Overlay journals and data publishing in the meteorological sciences. *Ariadne*, 60. Available at http://www.ariadne.ac.uk/issue60/callaghan-et-al/ [Last accessed 11 August 2014].

Callaghan, S A, Waight, J, Agnew, J L, Walden, C J, Wrench, C L and Ventouras, S 2013 The GBS dataset: measurements of satellite site diversity at 20.7 GHz in the UK. *Geoscience Data Journal*, 1(1): 2–6. DOI: 10.1002/gdj3.2.

Chappell, P R and Lorrey, A M 2013 Identifying New Zealand, Southeast Australia, and Southwest Pacific Historical weather data sources using Ian Nicholson's log of logs. *Geoscience Data Journal*, 1(1): 49-60. DOI: 10.1002/gdj3.1.

European Commission 2013 *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, Version 1.0*, 11 December. Available at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf [Last accessed 11 August 2014].

Garcia-Herrera, R, Können, G P, Wheeler, D A, Prieto, M R, Jones, P D and Koek, F B 2005 CLIWOC: a climatological database for the world's oceans 1750–1854. *Climatic Change*, 73(1–2): 1–12. DOI: 10.1007/s10584-005-6952-6.

Guilyardi, E, Balaji, V, Lawrence, B, Callaghan, S, Deluca, C, Denvil, S, Lutenschlager, M, Morgan, M, Murphy, S and Taylor, K E 2013 Documenting Climate models and their simulations. *Bulletin of the American Meteorological Society*, 94(5): 623–627. DOI: 10.1175/BAMS-D-11-00035.1.

House of Commons Science and Technology Committee 2010 *Science and Technology Committee Eighth Report: The Disclosure of Climate Data from the Climatic Research Unit at the University of East Anglia*. Available at http://www.publications.parliament.uk/pa/cm200910/cmselect/cmsctech/387/38702.htm [Last accessed 11 August 2014].

National Science Foundation 2010 *NSF Data Sharing Policy.* Available at http://www.nsf.gov/bfa/dias/policy/dmp.jsp [Last accessed 11 August 2014].

Oliver, J and Kington, J A 1970 The usefulness of ships' log-books in the synoptic analysis of past climates. *Weather*, 25: 520–528. DOI: 10.1002/j.1477-8696.1970.tb04103.x.

Otung, I E, and Savvaris, A 2003 Observed frequency scaling of amplitude scintillation at 20, 40, and 50 GHz. *IEEE Transactions on Antennas and Propagation*, 51(12): 3259–3267. DOI: 10.1109/TAP.2003.820960.

Research Information Network 2008 *To Share or Not To Share: Publication and Quality Assurance of Research Data Outputs, Main Report.* Available at http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf [Last accessed 11 August 2014].

The Royal Society 2012 *Science as an Open Enterprise. The Royal Society Science Policy Centre Report.* Available at http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf [Last accessed 11 August 2014].

Research Councils UK 2013a *RCUK Policy on Open Access and Supporting Guidance.* Available at http://www.rcuk.ac.uk/documents/documents/RCUKOpenAccessPolicy.pdf [Last accessed 11 August 2014].

Research Councils UK 2013b *RCUK Policy on Open Access: Frequently Asked Questions.* Available at http://www.rcuk.ac.uk/RCUK-prod/assets/documents/documents/OpenaccessFAQs.pdf [Last accessed 11 August 2014].

Science and Technology Facilities Council, Chilbolton Facility for Atmospheric and Radio Research, [Callaghan, S A, Waight, J, Walden, C J, Agnew J and Ventouras, S] 2009a *GBS 20.7GHz Slant Path Radio Propagation Measurements, Sparsholt Site.* Available at http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dep_11902946270621452 [Last accessed 11 August 2014]. DOI: 10.5285/E8F43A51-0198-4323-A926-FE69225D57DD.

Science and Technology Facilities Council, Chilbolton Facility for Atmospheric and Radio Research, [Callaghan, S A, Waight, J, Walden, C J, Agnew J and Ventouras, S] 2009b *GBS 20.7GHz Slant Path Radio Propagation Measurements, Chilbolton Site*. Available at http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dep_11902119479621181 [Last accessed 11 August 2014]. DOI: 10.5285/639A3714-BC74-46A6-9026-64931F355E07.

Science and Technology Facilities Council, Chilbolton Facility for Atmospheric and Radio Research, [Callaghan, S A, Waight, J, Walden, C J, Agnew J and Ventouras, S] 2009c *GBS 20.7GHz Slant Path Radio Propagation Measurements, Dundee Site*. Available at http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__ACTIVITY_dc47dc7c-8880-11e1-9490-00163e251233 [Last accessed 11 August 2014]. DOI: 10.5285/db8d8981-1a51-4d6e-81c0-cced9b921390.

Science and Technology Facilities Council, Chilbolton Facility for Atmospheric and Radio Research, [Ventouras, S, Callaghan, S A and Wrench, C L] 2012a *ITALSAT Radio Propagation Measurement at 20GHz in the United Kingdom*. Available at http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__ACTIVITY_e6d8b012-a65d-11e1-94b7-00163e251233 [Last accessed 11 August 2014]. DOI: 10.5285/3158D138-D592-4045-ADE4-B76CF9F42129.

Science and Technology Facilities Council, Chilbolton Facility for Atmospheric and Radio Research, [Ventouras, S, Callaghan, S A and Wrench, C L] 2012b *ITALSAT Radio Propagation Measurement at 40GHz in the United Kingdom*. Available at http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__ACTIVITY_52ad3c54-a663-11e1-ba03-00163e251233 [Last accessed 11 August 2014]. DOI: 10.5285/4A60EE2F-0FD1-4141-9244-7BEBF240BB49.

Science and Technology Facilities Council, Chilbolton Facility for Atmospheric and Radio Research, [Ventouras, S, Callaghan, S A and Wrench, C L] 2012c *ITALSAT Radio Propagation Measurement at 50GHz in the United Kingdom*. Available at http://badc.nerc.ac.uk/view/badc.nerc.

ac.uk__ATOM__ACTIVITY_f2984bd6-a664-11e1-ac44-00163e251233 [Last accessed 11 August 2014]. DOI: 10.5285/597C906A-B09E-4822-8B60-3B53EA8FC57F.

US Geological Survey 2000 *USGS Fact Sheet-036-00 March 2000*. Available at http://pubs.usgs.gov/fs/2000/fs036-00/fs036-00.pdf.

Ventouras, S, Callaghan, S A and Wrench, C L 2006 Long-term statistics of tropospheric attenuation from the Ka/U band ITALSAT satellite experiment in the United Kingdom. *Radio Science*, 41: RS2007. DOI:10.1029/2005RS003252.